DDF-HO: Hand-Held Object Reconstruction via Conditional Directed Distance Field

[1]Chenyangguang (Cyrus) Zhang* [2]Yan Di* [1]Ruida (Rudy) Zhang* [2]Guangyao Zhai
[3]Fabian Manhardt [1]Xiangyang Ji [2,3]Federico Tombari
[1]Tsinghua University [2]Technical University of Munich [3]Google

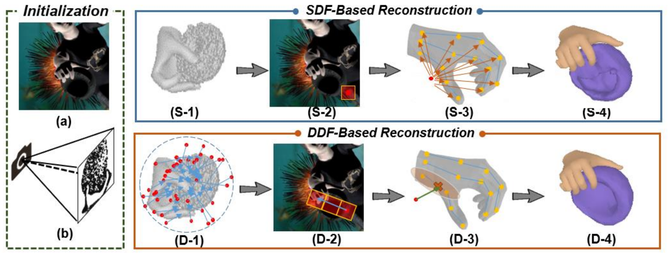
NEURAL INFORMATION PROCESSING SYSTEMS

Task Definition

[Hand-held Object Reconstruction]

- Given a single RGB image, DDF-HO predicts a 3D model for the object grasped by the hand. It is an essential technique with many practical applications, e.g. robotics, augmented and virtual reality, medical imaging.

Motivation

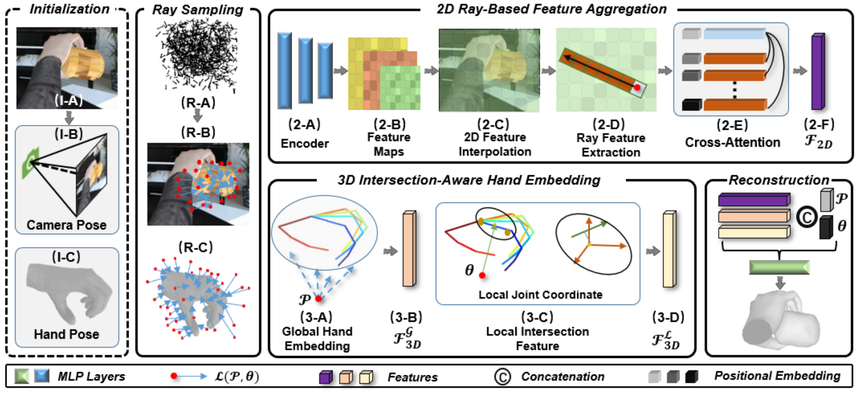


- **[Current Setting]** Most existing methods face challenges due to the use of **Signed Distance Fields (SDF)** as the primary shape representation. It is difficult to determine the nearest point on the object's surface to a sampled point without object shape prior. [1] addresses this challenge by aggregating features within a local patch centered around the projection of the point. However, this approach is unreliable when the sampled point is far from the object surface

- **[Our method]** We present **DDF-HO**, a novel **Directed Distance Field (DDF)** based Hand-held Object reconstruction. In contrast to SDF, DDF maps a ray, comprising an origin and a direction, in 3D space to corresponding DDF values, including a binary visibility signal and a scalar distance value measuring the distance from origin to target along the sampled direction. DDF offers advantages over SDF by providing better modeling of hand-object interactions. For each sampled ray, we collect its by combining 2D-3D geometric features via our 2D ray-based feature aggregation and 3D intersection-aware hand pose embedding.

Method

DDF-HO Pipeline



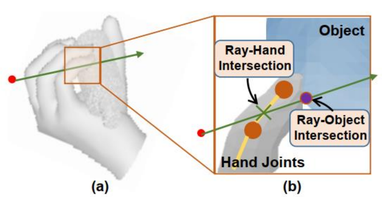
2D Ray-based Feature Aggregation

We project the sampled ray onto the image, yielding a 2D ray or a dot (degeneration case), and then aggregate features of all the pixels along the 2D ray as the 2D features, which encapsulate 2D local hand-object cues. we sample K_l points on the 2D ray, and extract local patch features $\mathcal{F}^{\ell}_{2\mathcal{D}}$ for all K_l points as well as the feature $\mathcal{F}^{p}_{2\mathcal{D}}$ of origin projection. Finally, we leverage the cross-attention mechanism to aggregate 2D ray feature

$$\mathcal{F}_{2D} = \mathcal{F}^{p}_{2D} + MultiH(\mathcal{F}^{p}_{2D}, \mathcal{F}^{l}_{2D}, \mathcal{F}^{l}_{2D})$$

3D Intersection-Aware Hand Embedding

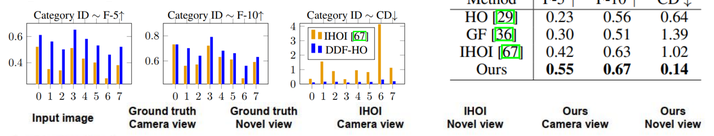
First, we calculate the shortest path from ray towards the hand skeleton, yielding starting point on the ray P_S and endpoint on the hand skeleton P_D. Then we detect K_{3D} nearest neighboring hand joints of P_D on the hand skeleton, using geodesic distance. Finally, P_S is transformed to the local coordinates of detected hand joints (3-C) and thereby obtaining $\mathcal{F}^{\ell}_{3\mathcal{D}}$ by concatenating all these local coordinates of P_S.



Experiments

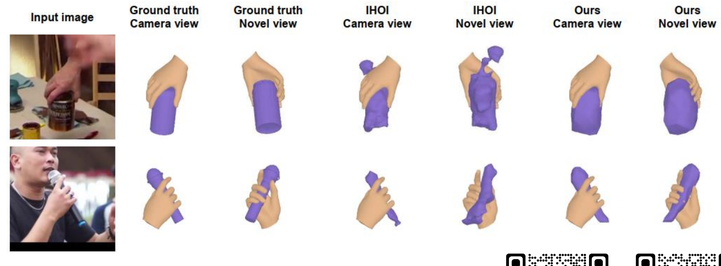
Main Results

- **[Obman]**



Method	F-5 ↑	F-10 ↑	CD ↓
HO [29]	0.23	0.56	0.64
GF [36]	0.30	0.51	1.39
IHOI [67]	0.42	0.63	1.02
Ours	**0.55**	**0.67**	**0.14**

- **[MOW]**



Method	F-5 ↑	F-10 ↑	CD ↓	F-5 ↑	F-10 ↑	CD ↓
IHOI [67]	0.10	0.19	7.83	0.09	0.17	8.43
Ours	**0.16**	**0.22**	**1.59**	**0.14**	**0.19**	**1.89**

- **[HO3D]**

Method	F-5 ↑	F-10 ↑	CD ↓	F-5 ↑	F-10 ↑	CD ↓
HO [29]	0.08	0.19	4.60	0.05	0.14	6.03
GF [36]	0.09	0.21	5.23	0.07	0.16	6.25
IHOI [67]	0.21	0.38	1.99	0.17	0.31	4.17
Ours	**0.28**	**0.42**	**0.55**	**0.24**	**0.36**	**0.73**

References

[1] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3895–3905, 2022.

[2] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11807–11816, 2019.

[3] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020

Code Paper Cyrus Rudy